E/ECA/PSD.4/62
16 January 1976
Original: ENGLISH ONLY

UNITED NATIONS

ECONOMIC AND SOCIAL COUNCIL

ECONOMIC COMMISSION FOR AFRICA

Fourth session of the Joint Conference of African Planners, Statisticians and Demographers

Addis Ababa, Ethiopia, 3-12 March 1976

REPORT OF THE SEMINAR ON STATISTICAL DATA BASE DEVELOPMENT

(For English-speaking African countries)

Addis Ababa, Ethiopia, 30 September - 4 October 1975

TABLE OF CONTENTS

	Paragraph
I. ATTENDANCE AND ORGANIZATION OF WORK	
(i) Opening and duration of the meeting	1 - 2
(ii) Attendance	3
(iii) Agenda	4
II. ACCOUNT OF PROCEEDINGS	
Electronic data processing requirements for national statistical systems (agenda item 2)	5 - 8
Co-operation in Europe in the field of statistical data base management (agenda item 3)	9 -18
Elements of the conception of a statistical data base (agenda item 4)	19 -23
Different data base approaches (agenda item 5)	24 -31
Structural analysis of statistical data base (agenda item 6)	32 -36
Economic Commission for Africa statistical data base (agenda item 7)	37 -44
The Swedish experience in data base management (agenda item 8)	45 -50
Statistical data base management "Eurostat experience and recommendation" (agenda item 9)	51- 58
Meta base documentation (agenda item 10)	59 -67

I. ATTENDANCE AND ORGANIZATION OF WORK

(i) Opening and duration of the Meeting

1. The Seminar on Statistical Data Base Development for English-speaking countries was held at ECA Headquarters, Addis Ababa, from 30 September to 4 October 1985. It was jointly organized by the Pan African Documentation System (PADIS) and the Statistics Division of the United Nations Economic Commission for Africa (ECA). The European Economic Community (EEC), Statistics Sweden and the United Nations Economic Commission for Europe (ECE) contributed to the meeting through the provision of lecturers and teaching material.

2. The Seminar was opened together with the workshop on National Accounts (30 September - 11 October 1985) by the Officer-in-Charge of the United Nations Economic Commission for Africa on behalf of the Executive Secretary.

(ii) Attendance

3. The Seminar was attended by five Electronic Data Processing (EDP) specialists responsible for EDP statistical processing from the following countries: Botswana, Ethiopia, Sierra Leone, Zimbabwe and by a representative of the League of Arab States.

(iii) Agenda

4. The following is the approved agenda for the workshop.

1. Opening address;
2. Electronic Data Processing (EDP) requirements for national statistical system;
3. Co-operation in Europe in the field of Statistical Data Base Management;
4. Elements of the Conception of a Statistical Data Base;
5. Different data base approaches;
6. Structural analysis of statistical data base;
7. Implementation of statistical data base "ECA" case study;
8. Steps in data base design "Statistics Sweden Experience";
9. Statistical data base management "Eurostat experience and recommendation";
10. Meta base (documentation)

II. ACCOUNT OF PROCEEDINGS

Electronic data processing requirements for national statistical systems (agenda item 2)

5. This paper was presented by a representative of the secretariat. It was emphasized that the question of Electronic Data Processing (EDP) is becoming quite important in many African countries. For example, statistics describing many aspects of the population and business environment are essential to central government. Such statistical information include very many millions of data items, and therefore require that organized electronic data processing facilities are made use of.

6. It was further emphasized that in developing EDP facilities care must be given to the selection and instalation of appropriate hardware and the adoption of relevant software facilities. Emphasis should be given to widely used software packages. It was emphasized that statistical data are information which is hopefully accurate, but it is their interpretation or conversion to knowledge or intelligence which is important to policy making, planning and government administration, and therefore protection should be given both from loss for economic reasons and also for confidentiality and privacy reasons.

7. Discussions of the paper centred around the approaches to systematically organized EDP work at national level. Some participants felt that EDP facilities that exist at national level have not been systematically established and therefore do not meet the standards as outlined in the paper.

8. Participants were requested to briefly give the status of EDP facilities in their countries. During the presentations, it was observed that quite a number of countries represented in the workshop had problems of staff, software and hardware. The question of how to place EDP facilities in the whole context of statistical work in a country also came up. Many felt that EDP is at different stages in different countries, and the improvement of EDP facilities is dependent on the growth of national statistical offices. Generally, most participants felt that there is a great need for properly managed EDP facilities at national level.

Co-operation in Europe in the field of Statistical Data Base Management (agenda item 3)

9. Mr. Christer Arvas of Statistics Sweden who acted as a consultant from the ECE gave a presentation of co-operation in Europe in the field of statistical data base management.

10. He pointed out that there were two reasons for putting the item on the agenda. In the first place, it was naturally hoped that results from the European work would be possible to use in Africa. Secondly, he pointed out that it might be possible to apply, in a modified form, some of the methods of work applied in Europe also in the African region. His presentation covered mainly the work carried out within the Statistical Computing Project (SCP) of the ECE and UNDP. In addition to the documentation already provided by ECA to the participants of the seminar, he promised to send two documents: Rapid Guidelines and Base Operators Users Guide to the participants by mail.

11. He gave a background description of the work carried out by the SCP and how it had been organized. In addition to a steering committee with members from all 20 participating countries and four joint groups there had been a project co-ordinator stationed in Geneva and a technical advisor in Bratislava.

12. The work which was formally started in 1981 was carried out in one group for relational data base management, one group for data editing, one for tabulation and one for meta data questions. In the first group, the work had been concentrated on the Canadian system RAPID, in the second, on the Hungarian system AERO, in the third group, on development of a new system called INTERTAB and in the fourth group, on working out specifications for a meta data information system. In addition, a handbook on statistical data processing and a glossary had been produced by the SCP.

13. Mr. Arvas then went into more detail of the work carried out by the RAPID group. The RAPID guidelines produced by the group were meant to illustrate not only the RAPID system itself but also the principles behind a relational data base manager. The objective was that also countries where it would not be possible to run RAPID would be able to utilize the ideas in constructing their own systems.

14. Within the framework of the group, major implementations of RAPID had been monitored in Hungary, GDR and Yugoslavia. Sweden and Canada already had several statistical systems running under RAPID.

15. He then went into a more detailed description of the system and its different interfaces both to users and to other external systems such as SAS and SPSS for statistical analysis.

16. In addition to development of the RAPID software itself, the group had developed a number of base operators which could be applied to any set of relational files.

17. Finally, Mr. Arvas related some evaluation studies of RAPID which indicated that the system was very efficient in statistical applications but that one has to be careful with the file design to ensure efficient runtime operation.

18. In the following discussion, it was stressed that it was difficult to directly apply SCP software in Africa because the computers used were often

of a type on which the SCP software could not directly be used. Mr. Arvas stressed the fact that the whole SCP work had been sponsored to some extent by the UNDP and that of course all results available could be used by statistical offices in Africa free of charge.

Elements of the conception of a statistical data base (agenda item 4)

19. In the presentation of the background paper for this agenda item, it was stated that a statistical data base, the purpose of which is to facilitate storage, manipulation and processing of the whole statistical data produced in a country should, if it must be effective and useful, include such characteristics in its conceptual planning phase as delivering of timely statistical information to a large number of users, integration of statistics as far as concepts, definitions, classifications and methods are concerned and ensuring the chronological continuity of data and their storage in systematic manner as cumulative elements of a data resource.

20. The three stages in designing a data base were mentioned. The first two phases, namely the conceptual and logical implementation phases were thoroughly discussed. The third phase, the physical implementation phase was left aside for discussion under another item.

21. As far as the management of data base is concerned, emphasis was put on the role of the data base administrator(s). The administrator of a data base being the person or group of persons responsible for the management of the data base and its general control, it was proposed that the administrator be given a high post equal to that of a Head of Department in the functional organization of a statistical office, and be answerable to the Head of office.

22. In the discussions which followed, the profile of a data base administrator was discussed. It was stated that for a statistical data base it was not necessary for the administrator to be an Electronic Data Processing (EDP) specialist. He might be a statistician capable of dialoguing with EDP specialists or an EDP specialist with a strong background in statistics. As a matter of fact in practice, it is unlikely for one person to combine the technical expertise that is needed, knowledge of all offices/departments data, and negotiation skills. The functions of the data base administrator may therefore be carried out by a group of persons. An important factor is the fact that the data base administrator is directly responsible to end-users. He should, therefore, be able to provide valuable counselling and consulting services to data end-users.

23. The suggestion according to which the data base administrator should be answerable to the Head of statistical office was considered further in the discussions which followed the presentation by the representative of the secretariat. The opinion of most participants was that the data administrator should be answerable to the Head of the EDP Centre. It was also pointed out that the above suggestion was irrelevant for those national statistical offices which do not have their own EDP Centre.

Different data base approaches (agenda item 5)

24. The subject was introduced by a representative of the secretariat. He stated that the development of any data base requires in the first stage a description of the application envisaged in terms of data to be stored, and of the processing to be made. This description can be made only if a model of data is available for understanding and interpreting this application. The organization of data is represented by a model of data, that is the set of concepts and related rules enabling the description of data. The model of data is a schema according to which data are logically organized. It represents, therefore, a logical data organization approach.

25. The representative of the secretariat further described three main basic logical data models in terms of definitions, advantages and disadvantages. They were namely: the hierarchical or tree model, the plex or network model, and the relational model.

26. It was stated that comparisons between the three models were difficult to make. Hierarchical and plex models are dependant on both logical and physical data description levels whereas the relational schema is described without explicit reference to data representation or access. Therefore, any change occurring in the internal representation of data in the hierarchical or network models lead to the modification of programmes which accessed the data, although the information remains the same. Furthermore, obtaining answers to a query in these two models requires a very heavy formulation of this query almost always in terms of algorithms, which is difficult to grasp for a non-programmer. In the relational model, on the other hand, the internal representation of information can be selected and modified without constraints and without any interference with the formulation of requests. It is not even necessary to know this internal representation of data, which is an advantage for non-programmer users.

27. The representative of the secretariat concluded his introductory remarks by stating that no model is the best in all cases. In the choice of logical model of data organization, environmental conditions should be taken into consideration such as: volume of data and its growth, frequency updates, efficiency in data loading, ease of use, speed and efficiency in retrieving data;

preliminary awareness of the use of the data base, type of users (programmers, non-programmers), type of the real world structure, technical support and training facilities, type of desirable security, costs, etc. These numerous selection criteria may contradict each other. Compromises have always to be made in such a selection. The relational model seems to be more suitable for statistical processing environment in that statisticians mostly view data as tables but the questions of performance for the existing prototypes as well as of risk of semantic disintegrity (meaningful relations leading to erroneous information) were also pointed out. It was observed that most of the time there is no possibility of chosing a logical data model. It is imposed by the hardware and/or the DBMS at hand.

28. In the discussion which followed the presentation by the secretariat, it was generally agreed among those who have experienced the development of data bases that it was difficult to compare various logical data models, and therefore to advise which could be the best in all cases for statistical processing environment. However, for the Statistical Computing Project (SCP) the Data Base Management System (DBMS) preferred was the relational model. Thus, RAPID was selected. The seminar was informed that Eurostat of the European Economic Community did not choose logical data models for its data bases but made use of all of them according to cases. Indexed sequential file technique was used in CRONOS for handling time series, network model in IDMS for tables, and relational model in ADABAS for external trade statistics.

29. It was pointed out that the logical data models presented were the basic ones, and that the reality was more complex in that several versions exist for a given model in practice.

30. It was also stated that the fact that the way data are accessed was not required in the relational data model made it easier to facilitate and ensure greater data security.

31. It was recognized that there was a performance problem with the existing Data Base Management Systems based on the relational model. However, it was stated that this problem depends also on the way the data base is implemented, particularly in terms of data physical storage and access. For instance, both RAPID and RELIANCE (used in Zimbabwe) were based on the relational data model but did not have the same data physical storage and access.

Structural analysis of statistical data base (agenda item 6)

32. A representative of the secretariat presented document ECA/STAT/SD3D/SM/85/4 entitled "Structural Analysis of Statistical Data Base". The representative expressed the complexity of statistical data in that there are many fields of

statistics each having different attributes. The statistical data is needed in a variety of forms by different categories of users statisticians, demographers, economists, planners and policy makers, etc. to meet their analytical requirements.

33. In developing a statistical data base management system where the volume of data is large, fast and very flexible access for retrieval, economy of storage, convenient, fast and ease of update and the ability of the data base management system to keep file integrity are very crucial.

34. There are many different file organizations each having its own structure and access procedure. These basic file organizations and the access methods used are the building blocks of the different data base types.

35. The three different types of data bases, external representation, the internal structures and the different access methods and techniques that could be used to build a statistical data base were analysed. The paper also proposed some techniques for solving problems of statistical data bases.

36. In the discussions which followed the proposal for the network data base approach was criticized in that a statistical data is in effect a relation and that a relational data base could also handle the statistical data base. There was a was a lengthy discussion on the different approaches and their applications. It was agreed that the application of the statistical data base determined its structure. It was also agreed that an integrated statistical data base structure on all spheres of statistics suitable to meet the various needs cannot be formulated at the present seminar. A more detailed study and analysis of the statistical data base structure should be done and a similar seminar should be held to discuss the issue.

Economic Commission for Africa Statistical Data Base (agenda item 7)

37. A progress report on the Development of the ECA Statistical Data Base (document ECA/STAT/SD&D/85/5) was presented by a representative of the secretariat. The ECA Statistical Data Base is organized in three levels: Level I provides comprehensive country profiles.. It comprises data on population, employment, national accounts, agriculture, industry, transport and communications finance, international trade, education and health. Level II contains the bulk of the statistical data at very low levels of aggregation on magnetic tapes and discs. Level III contains other time series data. The implementation of level I depends on the progress of levels II and III. The work on level II is substantial and has been delayed due to the lack of resources, namely adequate computer time and facilities as well as staff.

38. The over-all system design of the ECA statistical data base has been completed and level III data base is being implemented. Over 4800 statistical series identifiers have been developed, coded and entered. Over 101,000 time series statistical data on 12 different sectors of statistics have been entered. Table generator package and programmes to process data from foreign tapes have also been developed.

39. The ECA Statistical Data Base structure, content and retrieval system was presented as a case study to show the problems that exist in building statistical data bases and possible solutions.

40. There was much discussion on the ECA Statistical Data Base design, its portability to member countries and the type of services the secretariat would give to member States regarding Statistical Data Base Development.

41. It was stated that although the ECA Statistical Data Base could be, in its present stage, transferred to national statistical offices with HP 3000 computer, resources were needed for adaptating it to national requirements, this adaptation being even more complex for other computer makes. A representative of the secretariat stated also that the secretariat could, as far as resources permitted, assist member countries which did not receive other technical assistance in the development of their data base.

42. To the question why data like population data which had specific structure were put in the form of time series, it was replied that the main reason was the ease for producing ECA statistical publications, that an approach was used to present all these data as time series in order to retrieve the data for the year needed. It was also recalled that the present ECA Statistical Data Base represented only the level III and that its contents were to be derived from the level II which was expected to contain all ECA Statistics Division's data at the very lowest level of aggregation. It was decided to start with level III on account of many difficulties and problems encountered in developing the level II.

43. A representative of the secretariat briefed the meeting on the PADIS project of which the ECA Statistical Data Base is one component. Thus, the ECA Statistical Data Base is also sometimes referred to as PADIS-STAT.

44. The relation of the ECA Statistical Data Base with international data bases was questioned. It was stated that the UNCTAD's Economic Time Series Data Base (ETS) guided the design of the ECA Statistical Data Base. In addition, inputs to the data base were provided by various international organizations who supplied data tapes to ECA.

The Swedish experience in data base management (agenda item 8)

45. Messrs. Christer Arvas and Lars Olsson described steps in data base design based on experiences from Statistics Sweden.

46. Mr. Olsson introduced the subject by giving examples of different applications in a national statistical office where data base management systems could be used. He distinguished between survey systems, register systems and output oriented systems (statistical data bases). These systems are distinct in the way data are being entered, processed and retrieved. The result is that there will be different requirements on the data base management system.

47. Mr. Arvas introduced a discussion about the functions of a data base administrator. He stressed the importance of clarifying the responsibilities for different aspects of data management.

48. The different steps of the systems design process used in Sweden were presented by Mr. Arvas. They were: Infological modelling, usage of flat files, and an extensive use of generalized programmes. Examples were given both from Statistics Sweden and the Central Statistical Office of Zimbabwe.

49. Finally, Mr. Olsson gave a presentation of the Sweden DBMS for statistical data, AXIS. He stressed the importance of clearly defined goals and user groups requirements in development of statistical data base management systems.

50. In the discussion which followed, it was stressed that the approach to be selected is dependent on the role of the statistical office and the resources available.

Statistical data base management "Eurostat experience and recommendation" (agenda item 9)

51. The representative of the Statistical Office of the European Communities (SOEC) indicated briefly its place within the Commission of the European Communities showing its links with other general directorates, and particularly with informatics. He further stated that the history of the data banks could not be disassociated from those of SOEC itself. This history could be roughly broken down in three broad periods: the periods of harmonization, data collection, and dissemination.

52. During the first period, SOEC could set rules and imposed them. A statistical system (the European system of accounts, ESA) was developed including concepts, nomenclatures (NIMEXE, NACE, etc.) as well as surveys (labour force, salary structures, agricultural structures, production, etc.). The manpower was mainly oriented towards methodologies, definitions, and even rights. If data bases were developed at this stage, SASLIM an automatic system of management of nomenclatures (200 nomenclatures and links between them) would have been typical of this first generation data base period.

53. In the second period, it was necessary to collect, store, and process data. Two types of concerns came into sight.

(a) Disseminate quickly the results of data collection for the needs of the Commission, hence the design and creation of CRONOS which was a data base aiming at storing and disseminating quickly a small amount of data (10 to 20 thousands time series), then create various data bases geared towards publications

(b) meet quickly the needs of different new economic, agriculture, foreign trade, and finance policies of the commission with a view to supporting these policies as well as to assessing them. There followed an extension for the development of a set of data bases all meeting internal needs related to the policy of the Commission, and the needs for publication CRONOS and COMEXT were data bases typical of this second generation.

54. This development in spite of a hardware and software policy which was sometimes uncertain and even contradictory led to a very structured organization:

(i) Each Electronic Data Processing project had to be classified based on whether it related to only one SOEC organizational unit or to several general directorates, or even to all organizational units;

(ii) To each level of classification of projects would correspond determined hardware capabilities

- 1 SOEC organizational unit = 1 monopost microcomputer
- More than 1 organizational unit = 1 multipost micto/mini computer
- More than 1 general directorate = mainframe

(iii) To each level of classification of projects would correspond also various responsibilities shared between the Data Base Administrator (who decides or better defines the storage structure and access strategy, defines authorization checks and validation procedures, defines a strategy for back-up and recovery, monitors performance and responds to changes in requirements, interfaces with Data Base Managements System vendors or developers), and the Data Base Manager (who is in charge of the information, the contents, the interfaces between statisticians and end users, statisticians and informaticians etc.). A well structured organization implies that each application be motivated and targetted for users as well as for products. negotiations between statisticians expressing their needs in order to achieve their objectives, and informaticians trying to determine the best way (i.e. purchase or in-house development) to satisfy requests.

55. During the third period, faced with a vast amount of data, one begins to wonder to whom they are useful. The result of investigations is that the investment made does not bear any relation to the use. Although it is not the aim of a statistical office to make profit, it is a public service and must, therefore, provide data required in the most rapid and appropriate form. Some of the causes of this small use of the data can be traced to the following:

(a) Data bases were developed little by little in parallel with statisticians' needs with the system available at that time, thus resulting in time in the incompatibility of access between specialized data bases;

(b) Different software or hardware;

(c) Lamentable lack of documentation (catalogue, definition);

(d) Insufficient harmonization;

(e) Narrowness of field of comparisons.

56. Most data bases were developed to meet statisticians' needs (publications or analysis), that is for experts who know how to define and search for what they need, ("Get principle").

57. Multipurpose data bases must, therefore, be avoided. Sets of data with end users as target must as much as possible be created, with search systems (key-work, hierarchical, etc.), catalogues, and perfectly related to publications where it will be possible to receive "information" (and not to get data) which can be pre-presented and pre-formated ("Bring principle"). The second generation data base (CRONOS) will of course remain for use by statisticians. It will be well interfaced for processing, analysis, and publications which will feed third generation data bases which will be designed by communication specialists (picture, text vendors, communication and videotex specialists, etc.).

58. This is an exciting challenge for the statistician. No longer will he be only a data producer but he will also become an information specialist.

Meta Base Documentation (agenda item 10)

59. No background document was prepared for this item. Instead, a representative of the secretariat made an oral presentation of meta data and meta information system.

60. He explained that each piece of statistical data has different names. Statistical data is huge in volume and has complex storage and processing requirements. The management and improvement of such data, its control, storage and dissemination have led to the evolution of a well defined and structured documentation about the data itself. Meta data is data about data and the system of documenting the information on meta data is referred to as meta information system.

61. After these introductory remarks, there was a general discussion on the underlying concepts and definitions of meta data and meta information system.

62. The experience of Statistical Computing Project (SCP) of METIS (META Information System) was described by the representative of the SCP from Statistics Sweden. He noted that everything needed to describe the value of a statistical data is meta data. And users need meta data to properly interprete the data. Also system constructors need technical information about the data. File and record descriptions are the next level of a data base used as meta data. The description of each parameter used to define statistical data need be catagorized, documented and well defined as part of meta data. The classification and storage of meta data in a form of catalogue is one way of building the meta information system. Data dictionary is a special catalogue of meta information system which is used to describe the attributes of the statistical data.

63. It was generally agreed that there does not exist established terms and concepts of meta data. They differ in meaning and application in different systems.

64. Meta data could have two approached. One is a separate system of a set of nomenclature descriptions. The other is to incorporate the meta information system in the data base. Some experience in building the meta information system independent of the data base has proved a failure. In national statistical offices, catalogues of the pertinent statistical parameters should be incorporated within the system. This, of course, depends on the person in charge of the data.

65. The EEC delegate also described the experience and use of meta data. The data dictionary of the parameters of the statistical data was built using 48 characters long description referred to as LABELS in three languages. He also noted that the statistical series was further elaborated as being good data, data estimated or projected which are good for statisticians and data if used are at the users risk and discretion. He also discussed the meta information system CADOS (Catalogue Document Statistics) as a user information system. It makes use of pointers and flags to associate the data value with the meta information. This will result in data bases being no longer built for, among others, producing publications but on the contrary, the contents of statistical publications including data on data (notes, footnotes, etc.) will be used to feed data bases.

66. User documentation is also part of the meta information system. Generally, it is the responsibility of the statisticians to consider the building and usage of meta information system. And users cannot make proper use of the statistical data without the meta data. It was also stated that users generally do not consult the meta data built around the data base or even notes and footnotes on the publication they consult.

67. It was also stated that the development of meta information system on an international statistical data base is a very difficult task because it requires a lot of resources. Statistical definitions, concepts, nomenclatures used vary from one country to another. Their harmonization and standardization require that all countries concerned contribute to this development. These resources being scarce in countries of the African region, it was suggested that a cost-benefit analysis be made in terms of comparison between the cost of development and the ultimate use of the system before such an undertaking.